SPRINGBOARD DATA SCIENCE CAPSTONE PROJECT

PREDICTING NEXT DAY RAIN WITH MACHINE LEARNING

presentation by Sangeeta Jayakar

Problem Statement

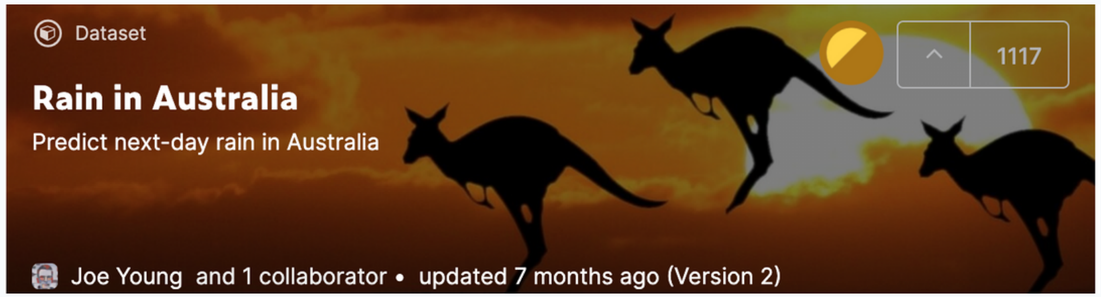
Can machine learning be used to predict next day rain?

Weather predictions have long been an important aspect of human civilization. Before modern science, people would attempt to forecast the rain by observing the clouds or the behavior of animals. The development of tools to measure atmospheric properties such as humidity, pressure, and temperature led to more advances in weather predictions, enhanced by the collection of global data and scientists who can interpret the data.

OUTLINE OF PROJECT

DATA WRANGLING	EXPLORATORY DATA ANALYSIS	PRE-PROCESSING	MODELING
choosing dataset and cleaning	finding trends and relationships in the data	preparing the data for modeling	testing four machine learning models

Data Wrangling



Choosing the dataset:

- Dataset was found on Kaggle.com
- Data were collected from Australia's Bureau of Meteorology
- Daily weather observations from 49 locations across Australia over a 10-year period.
- Raw data contained 145,460 rows across 23 columns

DATA CLEANING

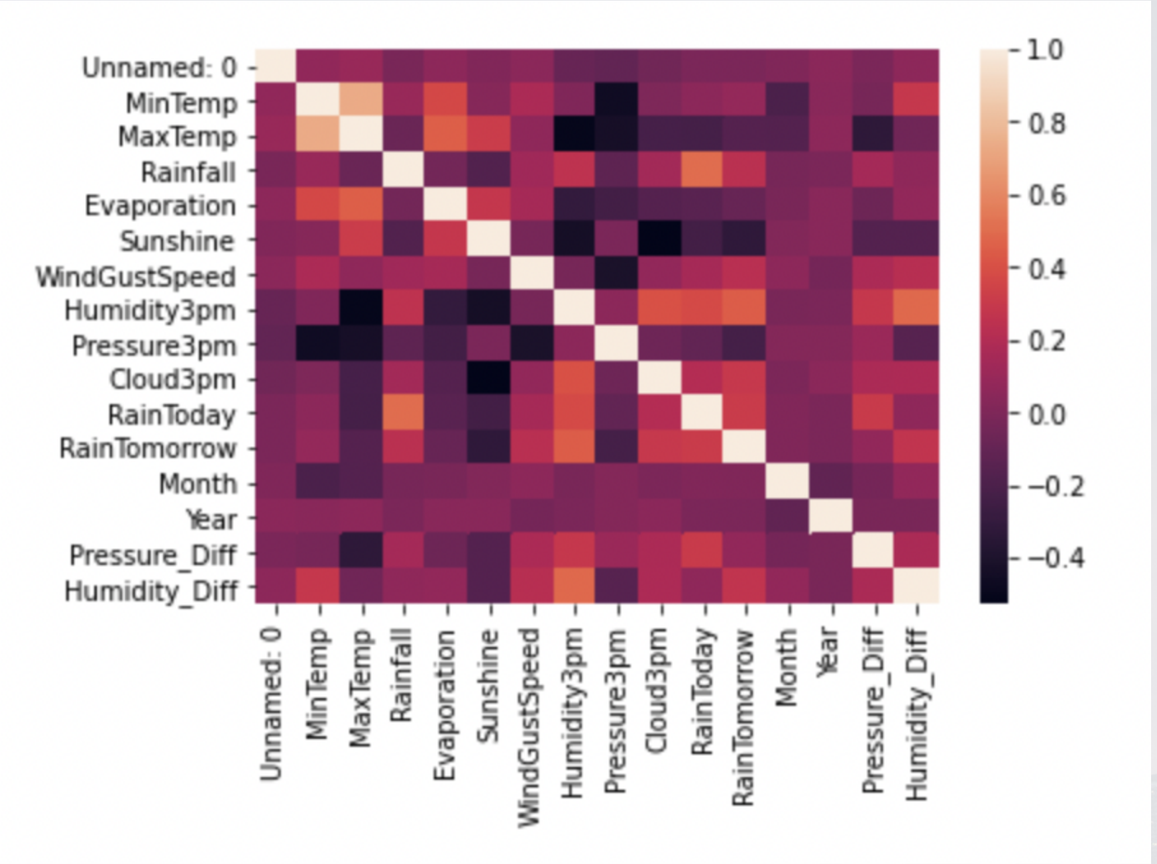
- **Dropped rows**
  rows with no entries for 'RainTomorrow' were dropped
- **Columns created**
  columns to calculate the difference in weather measurment from 9am to 3pm were created
- **Date columns**
  month names were extracted from dates and treated as categorical data
- **Columns dropped**
  columns for 9am measurements were dropped as they were redundant with 3pm measurements
- **Rain_Tomorrow**
  'Yes' and 'No' entries for Rain_Tomorrow and Rain_Today were replaced with numerica 1 and 0 to allow for ratio of rainy days to be calculated for EDA

Exploratory Data Analysis

CORRELATION BETWEEN FEATURES

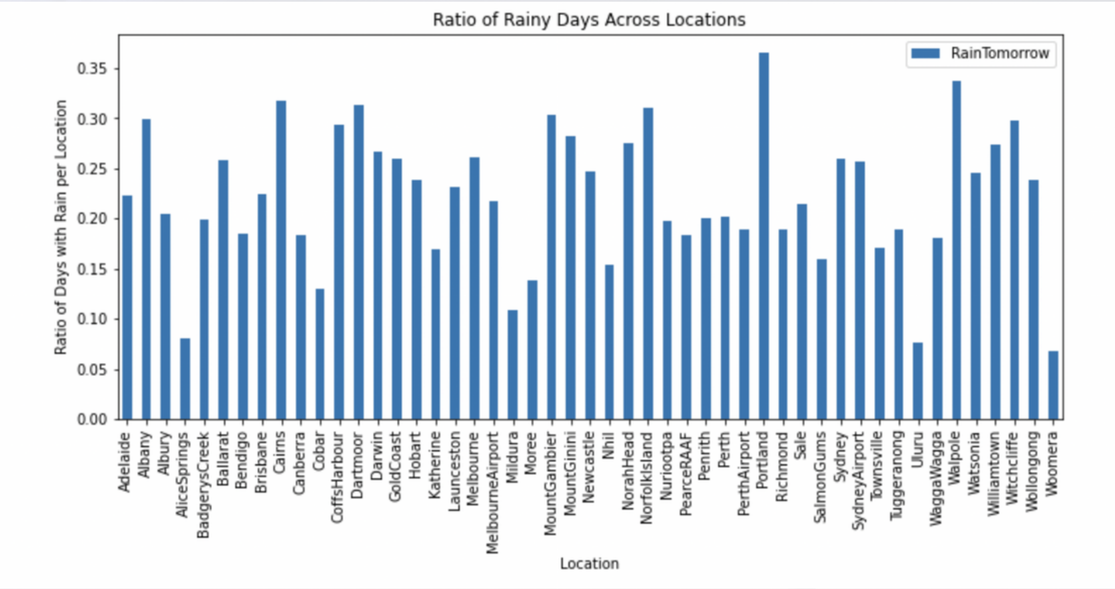
Target feature: 'RainTomorrow'

Features of interest: 'Humidity3pm' and 'Sunshine'



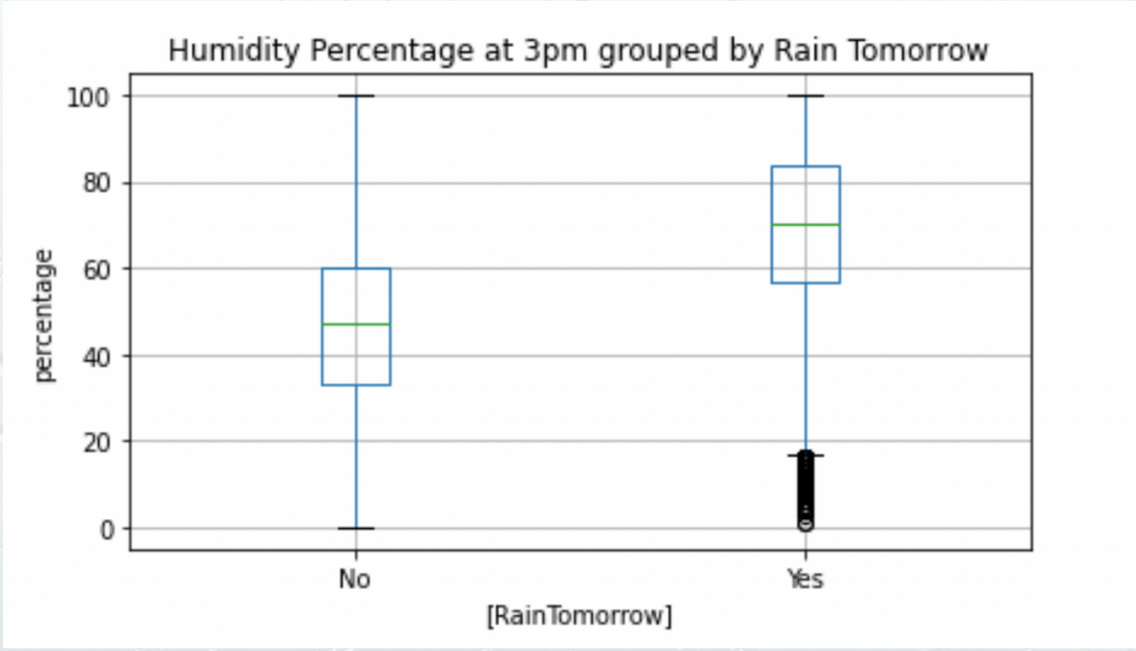
FEATURE: 'LOCATION'

Chi-square analysis shows location may be an important feature in predicting 'RainTomorrow'
p<0.001



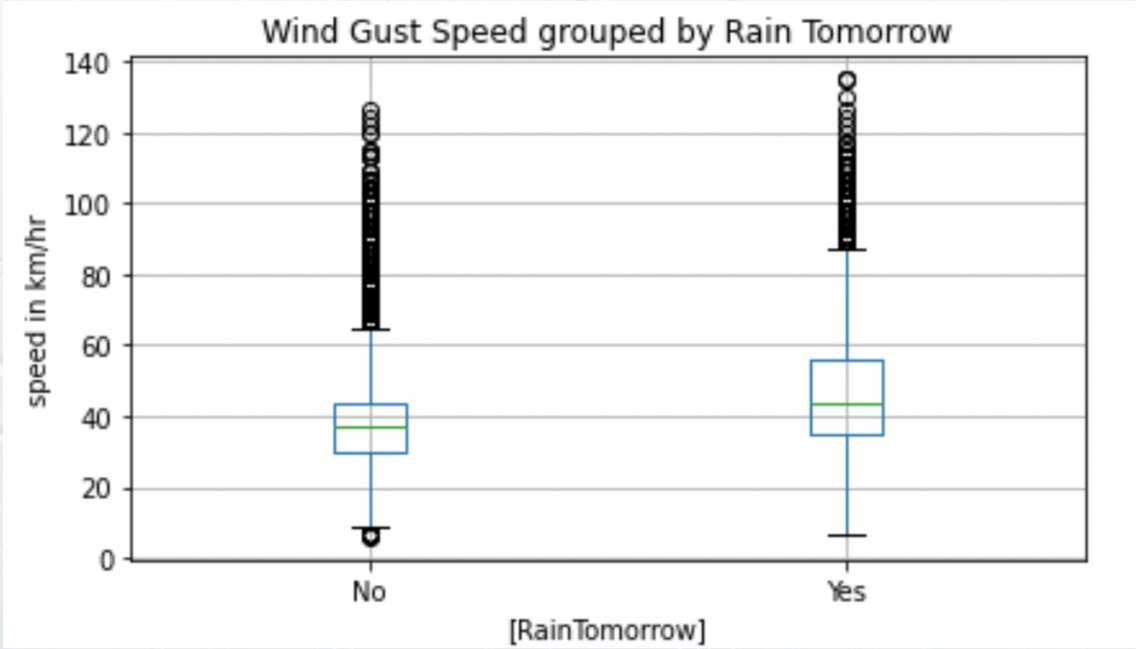
HUMIDITY AT 3PM AND RAIN TOMORROW

Humidity is significantly higher when it rains the next day (68.8%) compared to when there is no rain (46.5%), p<0.001 with t-test.



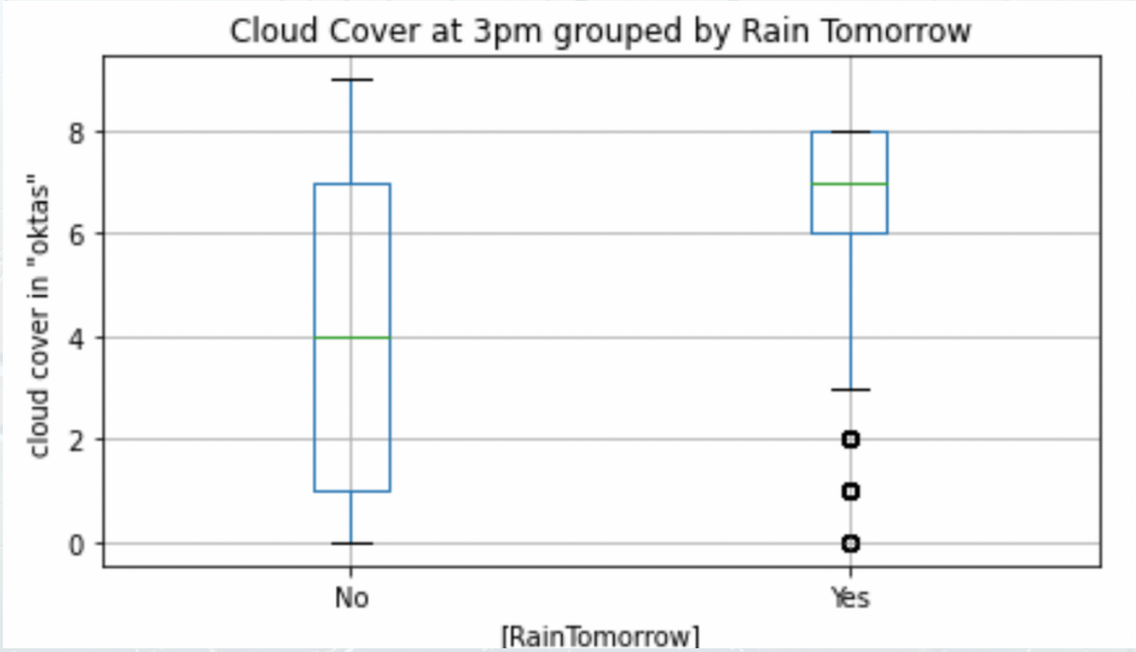
WIND GUST SPEED AND RAIN TOMORROW

The daily high wind gust speed is significantly higher when it rains the next day (45.9%) compared to when there is no rain (38.3%), p<0.001 with t-test.



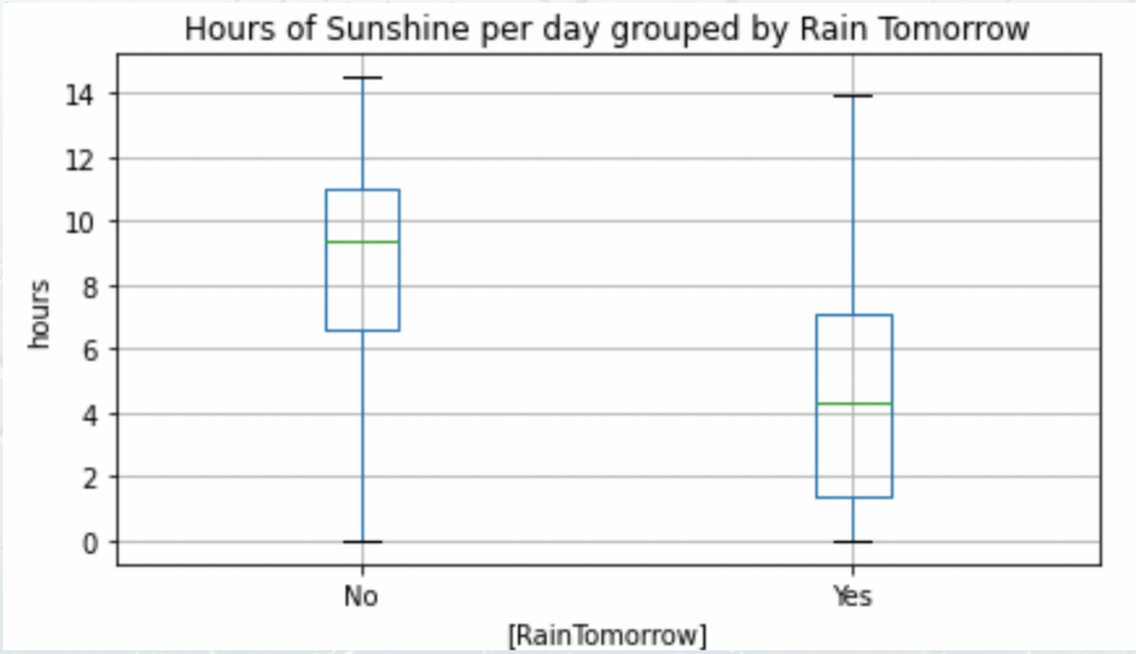
CLOUD 3PM AND RAIN TOMORROW

There is significantly more cloud cover when it rains the next day (7 oktas) compared to when there is no rain (4 oktas), p<0.001 with t-test.



SUNSHINE AND RAIN TOMORROW

Interestingly, the 'Sunshine' column had the opposite trend as the 'Cloud3pm'. When there is more cloud cover, there is less sunshine.



Pre-processing

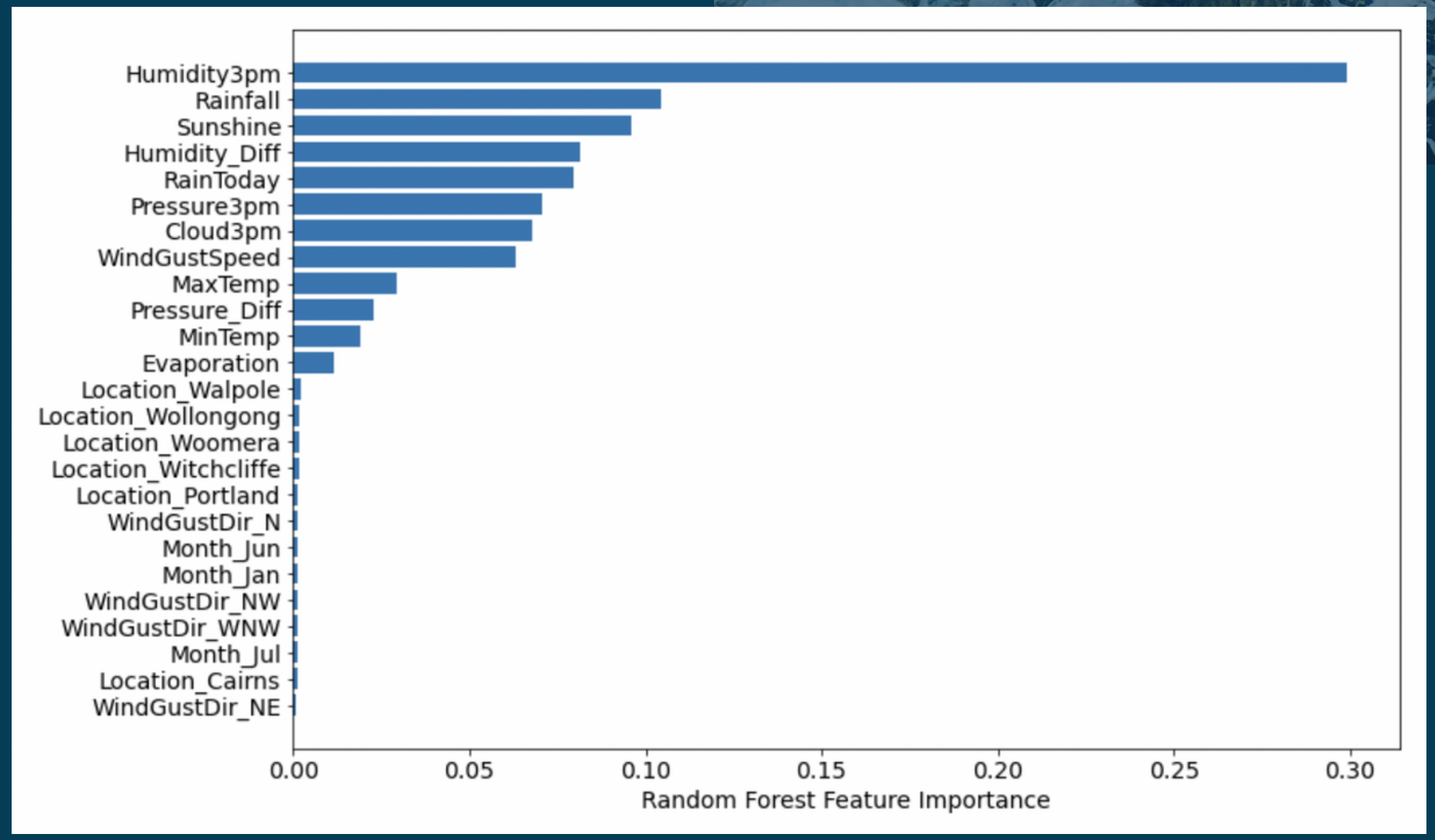
Prepare data for modeling:

- Imputed any remaining missing values with the 'median' of the column
- Created dummy variables for all categorical values
- Scaled the data using Scikit-learn's StandardScaler function

Modeling

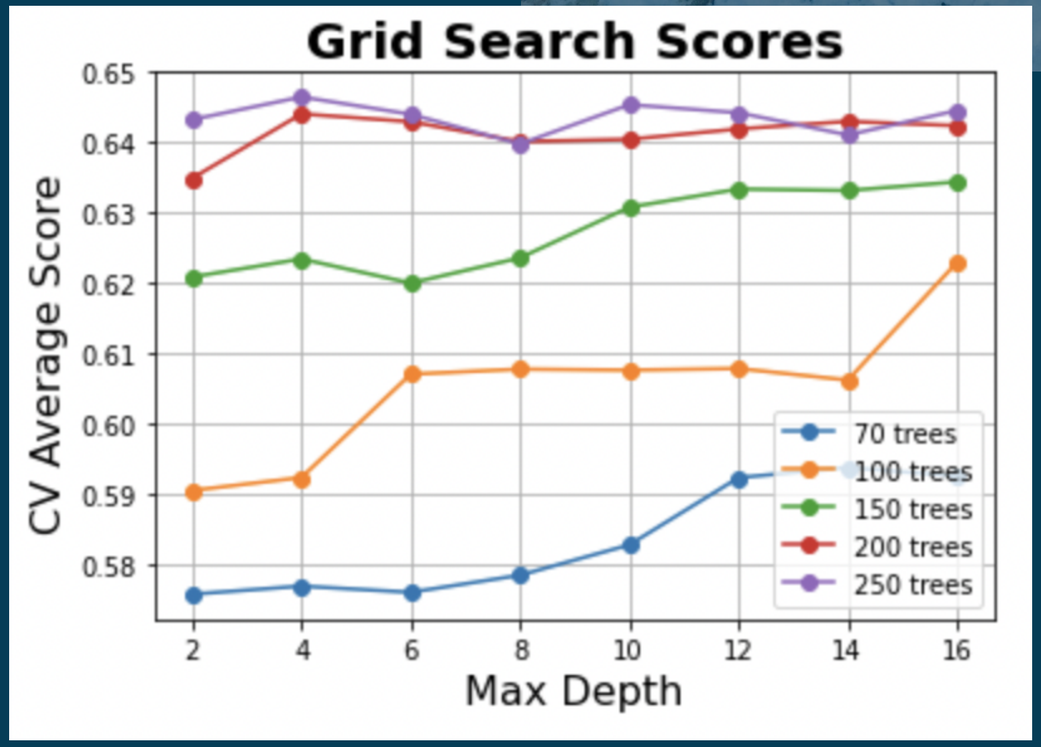
Random Forest

- Plotted the feature importances for RF model
- Some of the features we saw in the heatmap & EDA are showing up as highly important



Random Forest

- Used the GridSearchCV function to find optimal hyperparameters.
- Selecting 150 for number of trees and max_depth of 10.
- Using 'class weight balancing'



Model Performance

Model Name	F1 Score	Train Time	Best Hyperparameter Values
Decision Tree	0.611	0:00:00.841381'	{'max_depth': 6}
Random Forest	0.642	0:00:08.737272'	{'n_estimators': 150, 'max_depth': 10}
Gradient Boost	0.655	0:01:08.342800'	{'n_estimators': 200, 'max_depth': 6, 'learning_rate': 0.25}
Logistic Regression	0.620	0:00:00.852280'	{'C': 10}

Results & Findings

- Features that were important are intuitive to basic weather science: humidity, cloud cover, sunshine, air pressure, rainfall amount.
- Two features that I created during data wrangling, 'Pressure_Diff' and 'Humidity_Diff' also showed up in the feature importance, indicating that a drastic change in these measurements during the day are important for predicting rain.
- Interestingly, features that were not important: location and month. It seems like month would be important as there tend to be seasons where it rains more and seasons where there is drought. Location seems like it would matter because Australia has a diverse geographic climates (ie. tropical, temperate, and dessert/'outback')

CONCLUSION

After evaluating four models, the best model was the Random Forest. While the Gradient Boosting model had a slightly higher F1 score, the long training time may be a detterent for using this model, especially when there is a large amount of data. Random Forest gave a high F1 score with a relatively low training time. Also, the feature importances that were generated during the evaluation of the model are intuitive to the features that we explored in the EDA portion of this study.

Sources

Dataset:

Young, J. Rain in Australia, Version 2. Retrieved [March 2021] from [https://www.kaggle.com/jsphyg/weather-dataset-rattle-package]

Data source: http://www.bom.gov.au/climate/dwo/ and http://www.bom.gov.au/climate/data.

WEATHER FORECASTING, JOHN J CAHIR, BRITTANNICA.COM <https://www.britannica.com/science/weather-forecasting/History-of-weather-forecasting> [accessed July 2021]

THE BIRTH OF THE WEATHER FORECAST. BBC NEWS.COM <https://www.bbc.com/news/magazine-32483678>[accessed July 2021]

CLIMATE OF THE WORLD: AUSTRALIA <https://www.weatheronline.co.uk/reports/climate/Australia.htm> [accessed July 2021]